DOCUMENT RESUME

ED 161 376 HE 010 608

AUTHOR Overall, J. U.; And Others
TITLE Long-Term Stability of Students' Evaluations: A Longitudinal Study. AIR Forum Paper 1978.
PUB DATE May 78
NOTE 23p.; Paper presented at the annual Association for Institutional Research Forum (18th, Houston, Texas, May 21-25, 1978)
AVAILABLE FROM Dr. Overall, California State University, 1000 E. Victoria Street, Dominguez Hills, CA 90747

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Alumni; College Graduates; College Students; Comparative Analysis; *Course Evaluation; Effective Teaching; Followup Studies; *Graduate Surveys; Higher Education; Longitudinal Studies; Statistical Analysis; *Student Evaluation of Teacher Performance; Student Opinion; Student Reaction; *Time Perspective; Validity

ABSTRACT

The long-term stability of students' evaluations of instructional effectiveness was investigated. The focus was on the extent and nature of realtionships between end-of-term and retrospective ratings. Ratings given by alumni one year after program completion were compared with ratings given by these individuals in 85 undergraduate and graduate business administration classes, subsequent to course completion. Large and statistically significant stability coefficients were obtained between end-of-term and retrospective ratings, based on either individual or class average assessments. Mean differences were statistically significant for all but three of the 11 comparison items, but only one of these differences represented a shift of more than half a rating category. The results provide strong support for the stability of students' evaluations. (Author/SW)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED161376

THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper is part of the ERIC Collection of 1978 Association for Institutional Research Forum Papers. All papers in the Collection were recommended for inclusion by the AIR Forum Publications Review Committee.

Robert H. Fenske
Arizona State University
(Editor, AIR Forum Publications)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

HE 010 608

# Long-Term Stability of Students' Evaluations: A Longitudinal Study

J.U. Overall
California State University, Dominguez Hills

H.W. Marsh
University of Southern California

Robard Hughes
California State University, Dominguez Hills

H.L. Underbrink
California State University Dominguez Hills

This paper was the basis of a talk
presented at the

Annual Meeting of the ASSOCIATION FOR INSTITUTIONAL RESEARCH

Houston, May 1978

Requests for reprints should be directed to Dr. Overall,
California State University, 1000 E. Victoria Street
Dominguez Hills, CA. 90747.

ABSTRACT

The purpose of this study was to provide insight into the question of long-term stability of students' evaluations of instructional effectiveness. For, if students evaluate their instructors differently after having a period of time to reflect on class experiences, then the usefulness of such ratings for instructional, informational, or administrative decisions may have to be re-examined.

This study focused on the extent and nature of relationships between end-of-term (EOT) and retrospective ratings. Ratings given by alumni one year after program completion were compared with ratings given by these same individuals in 85 undergraduate and graduate business administration classes, subsequent to course completion.

Large and statistically significant stability coefficients (relative agreement) were obtained between EOT and retrospective ratings, based on either individual (median $r = .59$, $p < .01$) or class average assessments (median $r = .81$, $p < .01$). Mean differences (absolute agreement) were statistically significant for all but three of the eleven comparison items; however, only one of these differences represented a shift of more than half a rating category (1--9 scale).

These results provide strong support for the stability of students' evaluations. Students did not, to any significant extent, rate instructors differently in retrospect.

Long-Term Stability of Students' Evaluations: A Longitudinal Study[1]

End-of-term (EOT) assessments of instruction by students can provide important evaluative judgments about the effectiveness of the teaching/learning environment. A representative sampling of recent research has presented strong evidence that such EOT assessments are 1) reliable (Feldman, 1977), 2) valid (Centra, 1977; Marsh, Fleiner and Thomas, 1975; Marsh and Overall, 1978; Overall and Marsh, 1978), and 3) promote instructional improvement (Centra, 1972; Overall and Marsh, 1977; Pambookian, 1977).

However, critics of these positive findings suggest that were researchers to use follow-up ratings by graduates instead of EOT assessments by continuing students, the results might be quite different. In fact, some educators propose that because they allow students to develop perspective about and to obtain emotional distance from the person and situation being assessed, the ratings provided by graduates in retrospect are far more informative--from a course and program evaluation viewpoint--than EOT evaluations. If this is true, then the positive results of previous research concerning the validity and usefulness of students' evaluations--research that has relied almost totally on EOT assessments--may have to be re-evaluated.

This issue of the stability of students' evaluations is particularly important to the institutional researcher, for it is he (or she) who is often called upon to suggest, develop and

[1]The authors would like to acknowledge the extensive assistance in data collection and preparation of Ms. Jurie Imai.

periodically assess campus instructional evaluation programs. Given that EOT assessment is the most economical and commonly used form of instructional evaluation, it is essential to determine: 1) whether or not EOT and retrospective assessments differ, and 2) to what extent these differences--if any--have statistical or practical significance.

The purpose of the present research, then, was to investigate the correspondence between students' EOT and retrospective assessments of their courses and instructors. It examined a comparison of course and instructor ratings given by alumni one year after graduation, with assessments of the same courses and instructors just as these alumni were completing the course as students The extent of the correlation between EOT and retrospective assessments (relative agreement), and the amount of disparity between EOT and retrospective assessment item means (absolute agreement) were established to determine correspondence.

## Review of Research

Research on the stability of students' evaluations of instructional effectiveness can be characterized as either short-term or long-term. Short-term stability studies have compared ratings from one term to the next, or between mid-term and the end of the term. These studies have shown that over this relatively brief period of time, students' evaluations are highly stable (Costin, Greenough and Menges, 1971; Frey, 1976; Kohlan, 1973; and Overall, 1977).

Long-term stability studies might conveniently be defined as those concerned with both EOT evaluations and evaluations obtained at least one year after graduation. With an optimum research situation, such studies would compare ratings given by the same students both when they received the instruction to be evaluated, and later, after they had the opportunity to reflect on such instruction after graduation. If the span of time involved exceeded one quarter, only the most determined monitoring efforts would permit keeping up with these students after the course was over.

Given the difficulty of using the same students for both EOT and retrospective rating comparisons (essentially, a longitudinal design), previous researchers concerned with long-term stability have utilized a cross-sectional approach. In this procedure, students' ratings are compared with those of an alumni group assumed to be similar. For example, Drucker and Remmers (1951) compared alumni and students' average ratings of instructors, and found positive correlations between .40 and .68 on 10 teacher characteristics. These average ratings were similar for 6 of 10 items; alumni ratings were significantly lower ($p < .01$) on one other item (fairness in grading). While similar assessment items were used, instructor ratings were not based on the same courses, and alumni raters--by virtue of the cross-sectional design--were not identical with the student raters. Centra (1973) updated and expanded research in this area. Using a much broader data base, he compared rankings of 23 instructors

(based on students' average ratings on an overall summary item), with alumni rankings of instructors (based on a "best-worst" nomination scheme). The rank correlation for the student-alumni responses was .75 ($p<.05$). Marsh (1977) obtained similarly positive findings. In his study, however, students' and alumni ratings were based on the same class for each instructor considered; in the Centra study, no control was possible to insure that students' and alumni ratings of a particular instructor were based on his or her teaching of the same class.

Methodology

Subjects for the study were undergraduate and graduate students in business administration at a large California state university, who had completed academic requirements for their respective degree programs during academic years 1974-75, 1975-76, and 1976-77. A comparison of their EOT and retrospective assessments of courses and instructors provided the basis for this study.

Students' ratings of instructional effectiveness were obtained at the end of the term through use of the 28-item Student Assessment of Instruction (SAI) questionnaire (Overall and Marsh, 1974). This questionnaire permits students to assess instructional effectiveness with respect to five dimensions of the learning/teaching enrironment--Instructor SKILL, Course CHARACTERISTICS, Course STRUCTURE, Course VALUE/RELEVANCE, and Instructor-Student INTERACTION--using a 9-point rating scale (1 = very low/never to 9 = very high/always). Retrospective

instructor ratings were based on graduates' responses to an abbreviated 11-item SAI questionnaire. Those items loading highest on all but the course CHARACTERISTICS factor were included.

The EOT evaluation questionnaires were administered by the program director. At that time, students were asked to voluntarily place their identification numbers on the questionnaires, although they were specifically informed that this was not required. Students were informed that their evaluations would be used for both instructional improvement and program assessment purposes. One year after graduation, these students were sent the abbreviated questionnaire and were asked to re-evaluate their courses and instructors using the subset of eleven identical items.

The actual data analysis was performed on students evaluations of 85 classes. A subset of 10 individual EOT/retrospective evaluation combinations was chosen for each class; for those classes with more than 10 complete EOT/retrospective rating comparisons (original class size ranged from 18-32), 10 individual comparisons were randomly selected. Because results for both undergraduate and graduate students were similar, they were combined. Subsequent discussions of results in this paper will focus on this combination.

Long-term stability coefficients were computed using two separate procedures. In the first, EOT/retrospective assessments from the pre-selected group of undergraduate and graduate

respondants were correlated on an individual item-by-item basis. In the second, average instructor ratings from the classes were used. Pearson product-moment correlation coefficients were calculated to reflect EOT/retrospective assessment correspondence.

Mean differences based only on the class average samples described above were also calculated. A t test of the differences between means was computed to indicate level of statistical significance.

Results

*Stability Coefficients

The results shown in Table One indicate that in a relative sense, students' evaluations of instructional effectiveness are highly stable over time. A comparison of 850 pairs of course

-----Insert Table One About Here-----

and instructor ratings provided by individual students--both graduates and undergraduates--at the end of the term and after graduation, showed coefficients ranging from .46 to .68 ($p<.01$), with a median correlation of .59. These results indicate that when asked to re-evaluate their courses and instructors in retrospect, individual students were very consistent in their assessments.

Often, however, class average responses are of more practical use to the institutional researcher. Individual students' assessments are normally summarized by class and are provided to the faculty member for instructional improvement purposes. Thus, as described earlier, the present study also looked at the relationship between class average ratings at the end of the term and in retrospect. These results showed a similar but higher pattern of correlations ranging between .67 and .87 ($p<.01$), except for item 5 ($r = .38$, $p<.01$).[1] The median correlation using class averages was .81.

*Mean Differences

Mean differences between EOT and retrospective ratings are presented in Table Two. In three instances, retrospective

-----Insert Table Two About Here-----

ratings were somewhat higher than EOT ratings (items 2,3 and 10); in six instances, retrospective ratings were exceeded by EOT ratings (items 4,5,6,7,8 and 9). For two items (1 and 11), EOT/retrospective assessment differences were negligible. All rating means were above average (5.0 or higher on a 1-9 scale), indicating students' general satisfaction with their courses and instructors.

[1]This would be expected because each individual rating has a certain amount of random error associated with it; by using class averages, the random error tends to be cancelled out, producing relatively higher coefficients.

While eight of the 11 differences described above were statistically significant, only those items reflecting differences of at least .20 will be discussed further. Such differences represent average rating changes of at least one-fifth of a rating category, and probably indicate findings having the greatest practical significance.

Discussion

Stability coefficients obtained for the SKILL, STRUCTURE and INTERACTION evaluation dimensions were consistently high. These dimensions include items commonly found in most questionnaires used by post-secondary institutions for the evaluation of instruction. They reflect students' perceptions of their instructors' ability to clarify course objectives, present material in an organized manner, and stimulate interest and interaction. It is interesting to note that while students are called upon to assess relatively specific behaviour through these variables (items 1,2, 3,6.7 and 8), the stability coefficients were of similar magnitude as the coefficients for the more general summary items discussed below.

Looking at the overall rating variables, it is apparent that the overall course item (10) had the third highest stability coefficient in terms of individual assessments (.61) and the fifth highest coefficient (.84) in terms of class average assessment. The overall instructor item (9) had the highest stability coefficient in terms of individual assessments (.68) and the third highest coefficient (.86) in terms of class average assessment. These findings suggest that the magnitude of

ratings concerned with specific or general instructor or course related items are not--to any great extent--a function of time, and that students' ratings of their instructors are, in a relative sense, quite stable.

Finally, it is interesting to examine the two assessment variables identified as having the poorest stability over time, comparing either individual or class average ratings. These items reflected students' perceptions of the extent to which they would (did) pursue further study of the subject (item 5), and the extent to which the subject matter was (is) relevant to what they did then or would be doing (item 11). In both cases, students were asked to predict some future impact of the course, and to provide responses that were not based on easily observable criteria. Understandably, the stability of these assessments was lower than for assessments requiring less speculation. Thus, if educators use such items for evaluative purposes, they should be aware that EOT and follow-up assessments may differ to a greater extent than assessments concerned with items that are more directly observable.

With respect to mean differences, only one of the six assessment variables making up the rating dimensions of SKILL, STRUCTURE and INTERACTION (items 1,2,3,6,7 and 8) reflected a change greater than .20. This variable (item 6) was concerned with the extent to which students perceived their instructors to be considerate of their feelings and viewpoints. Generally speaking, then, mean differences on these dimensions were not significant.

Turning to the general rating items, some mixed results were obtained. While on the one hand, students' follow-up assessments of their instructors were significantly lower, the follow-up assessments of their classes were significantly higher. The reason for this contrast in findings cannot be determined from the data available; however, it is probable that these results in part reflect that the summary rating items may be based on elements not completely assessed through the other rating factors. In addition, summary ratings also may reflect individual weightings of the relative importance of other assessment factors, such that emphasis on each rating dimension assumed to reflect one aspect of effective teaching may receive different emphasis from individual to individual. Thus, the fact that mean differences for summary items were not completely consistent with mean differences for the SKILL, STRUCTURE and INTERACTION dimensions should not be surprising.

Looking at mean differences for the VALUE/RELEVANCE dimension, it appears that in retrospect, students perceived that they learned less of value from their classes (item 4), and that their actual study of the course material further after the class had ended (item 5) was less than expected. These results should be viewed in light of students' feelings about course relevance (item 11), where there was essentially no difference between EOT and follow-up assessments. Apparently, while students' perceptions of subject matter relevance did not change, their original perceptions that the subject material

would be both valuable and worthy of further study did. One might interpret these results as reflecting a program weakness, or merely the inability of students to accurately predict future value or study plans.

In summary, follow-up assessments tended to differ from those made at the end of the term, with most being lower in retrospect. While 8 of 11 differences were statistically significant, only five variables reflected shifts greater than .20, and only one variable (item 5) experienced a shift of more than half a rating category (from 5.83 to 5.28). All assessment means remained above average (5.0).

## Conclusion

In the present study, the major concern was the extent--if any--that EOT/retrospective assessments might differ. For if the differences were significant, then the use of EOT ratings in lieu of, or in the absence of, retrospective ratings might well be open to challenge.

The results of this study provide strong support for the assumed stability of students' evaluations of their courses and instructors. The large and statistically significant correlations obtained between EOT and retrospective ratings suggest that, after a period of time for reflection, students do not change their initial evaluative judgments, at least in a relative sense. Furthermore, while statistically significant mean differences were found between EOT and follow-up assessments on 8 of 11 items, only one of these differences reflected a shift of more

than one-half a rating category. Thus, in an absolute sense, the differences were of little practical significance.

At the annual meeting of a nationally recognized body of educational researchers earlier this year, one complete session was devoted to a discussion of what inputs from students might be useful for departmental and program evaluation purposes. Strong arguments were made for the use of general students' satisfaction (Braskamp, 1978) and for students' achievement (Centra, 1978; see also Hartnett and Centra, 1977). Given that, as outlined earlier, recent research has found students' assessments of their instructors to be reliable, useful, valid, and--as indicated in the present study--stable, then a good argument can also be made for their use in similar individual and organizational assessment activities.

In the short run, a great deal of evidence already exists to support the use of students' evaluations for formative assessment purposes. Individual instructors can use such feedback as a periodic check on their classroom effectiveness, while larger organizational units like departments can use this feedback as one indication of overall effectiveness at a certain point in time.

Students' evaluations also provide functional information for longer range, summative assessments. They indicate that neither maturity, situational considerations, nor the opportunity for emotional distance from the instructors or courses being evaluated, appear to alter such assessments. Thus, instructors and departmental units may use a collection of these assessments over time as one indication of effectiveness. In doing so, they

may be confident that students' assessments of their courses and instructors, whether obtained at the end of the term or in retrospect, will be quite stable.

BIBLIOGRAPHY

Braskamp, L.A. Student satisfaction as a useful criterion in departmental review. Paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Centra, J.A. The relationship between student and alumni ratings of teachers. Student Instructional Report No. 3. Princeton, N.J.: Educational Testing Service, 1973.

Centra, J.A. Should evidence of student learning be used to judge a department's effectiveness? Paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Centra, J.A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Centra, J.A. Two studies on the utility of student ratings for instructional improvement. Student Instructional Report No. 2. Princeton, N.J.: Educational Testing Service, 1972.

Costin, F., Greenough, W.T., and Menges, R.J. Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 1971, 41, 511-535.

Drucker, A.J. and Remmers, H.H. Do alumni and students differ in their attitudes toward instructors? Journal of Educational Psychology, 1951, 42, 129-143.

Feldman, K.A. Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education, 1977, 6, 223-274.

Frey, P.W. Validity of student instructional ratings: Does timing matter? Journal of Higher Education, 1976, 47, 327-336.

Hartnett, R.T. and Centra, J.A. The effects of academic departments on student learning. Paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Kohlan, R.G. A comparison of faculty evaluations early and late in the course. Journal of Higher Education, 1973, 44, 587-595.

Marsh, H.W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14, 441-447.

Marsh, H.W., Fleiner, H. and Thomas, C.S. Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 1975, 67, 833-839.

Marsh, H.W. and Overall, J.U. The validity of students' evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Paper presented at the meeting of the Association for Institutional Research, Houston, May 1978.

Overall, J.U. Students' Evaluations of Instructional Effectiveness: Validity and Utility. Unpublished doctoral dissertation, University of California, Los Angeles, 1977.

Overall, J.U. and Marsh, H.W. Cognitive and affective outcomes: Their relationship to effective teaching and students' evaluations of instruction. Paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Overall, J.U. and Marsh, H.W. The relationship between students' evaluations of faculty and instructional improvement. Paper presented at the Third International Conference on Improving University Teaching, Newcastle-upon-Tyne, England, June 1977.

Overall, J.U. and Marsh, H.W. Students' Assessment of Instruction (SAI): An Evaluation Questionnaire for the Adult Student. Technical Report No. 1. University of California, Los Angeles, 1974.

Pambookian, H.S. Feedback to instructors on their teaching to improve instruction. Paper presented at the Third International Conference on Improving University Teaching, Newcastle-upon-Tyne, England, June 1977.

TABLE ONE

Long-Term Stability Coefficients Based on Individual Students' Ratings and Class Average Ratings of Instruction

| | Rating Item | Combined Undergraduate/Graduate End-of-Term/Retrospective Rating Correlations | |
|---|---|---|---|
| | | N = 850 Individual Ratings[1] | N = 85 Class Average Ratings[2] |
| 1 | Purpose of class assignments made clear? | .55** | .76** |
| 2 | Course objectives adequately outlined? | .57** | .81** |
| 3 | Class presentations prepared/organized? | .61** | .79** |
| 4 | You learned something of value from class? | .56** | .82** |
| 5 | Did you study the subject further after the class was over? | .47** | .38** |
| 6 | Instructor considerate of feelings/views | .62** | .85** |
| 7 | Instructor involved you in discussions? | .59** | .87** |
| 8 | Instructor stimulated interest in subject? | .60** | .86** |
| 9 | Overall rating of this instructor? | .68** | .86** |
| 10 | Overall rating of this class? | .61** | .84** |
| 11 | Extent course material relevant to what you do now or will do in the future? | .46** | .67** |

**$p < .01$

[1] results based on 850 individual sets of students' evaluations of courses/instructors

[2] results based on class average ratings from 85 classes: averages derived from 10 ratings per class, with the 10 raters randomly selected if more than 10 ratings per class available.

TABLE TWO

Mean Differences between End-of-Term and Retrospective Ratings[1]

| | Rating Item | Combined Undergraduate/Graduate End-of-Term(EOT)/Retrospective(Retro) Mean Differences | | | |
|---|---|---|---|---|---|
| | | EOT Rating | Retro Rating | Retro/EOT Difference | Statistical Significance $t(84)$ = |
| 1 | Purpose of class assignments made clear? | 6.61 | 6.64 | + .03 | + .30 ns |
| 2 | Course objectives adequately outlined? | 6.45 | 6.62 | + .17 | +2.02 * |
| 3 | Class presentations prepared/organized? | 6.50 | 6.65 | + .15 | +1.53 ns |
| 4 | You learned something of value from class? | 6.88 | 6.66 | - .22 | -2.99 ** |
| 5 | Did you study the subject further after the class was over? | 5.83 | 5.28 | - .55 | -4.40 ** |
| 6 | Instructor considerate of feelings/views? | 6.89 | 6.53 | - .36 | -4.60 ** |
| 7 | Instructor involved you in discussions? | 6.74 | 6.58 | - .16 | -2.56 * |
| 8 | Instructor stimulated interest in subject? | 6.49 | 6.32 | - .17 | -2.15 * |
| 9 | Overall rating of this instructor? | 6.70 | 6.48 | - .22 | -2.65 * |
| 10 | Overall rating of this course? | 6.45 | 6.68 | + .23 | +3.54 ** |
| 11 | Extent course material relevant to what you do now or will do in the future? | 6.15 | 6.11 | - .04 | - .30 ns |

ns = not statistically significant
*$p < .05$
**$p < .01$
[1]results based on class average ratings from 85 classes; averages derived from 10 ratings per class. with the 10 raters randomly selected if more than 10 ratings per class available.